# Analyzing Reusability of Speech Corpus based on Statistical Multidimensional Scaling Method

*Goshu Nagino and Makoto Shozakai*

Information Technology Laboratory, Asahi Kasei Corporation
Atsugi AXT Main Tower 22F, Okada 3050, Atsugi, Kanagawa, 243-0021, Japan
{g-nagino, makoto}@ljk.ag.asahi-kasei.co.jp

## Abstract

In order to develop a target speech recognition system with less cost of time and money, reusability of existing speech corpora is becoming one of the most important issues. This paper proposes a new technique of applying a statistical multidimensional scaling method to analyze the reusability of a speech corpus. In the experiment using six speech corpora, which contains isolated words and short sentences used in car navigation system, an effect of the proposed method is evaluated by a usual approach of cross task recognition. Furthermore, the relationship among those speech corpora is clearly shown by the proposed method.

**Index Terms**: statistical MDS, task dependency, acoustic model

## 1. Introduction

Recognition accuracy is still extremely sensitive to the environmental conditions such as speaker characteristics, speaking style, background noise and domain. These issues are called task dependency. Task dependency has strong impact on recognition performance of Automatic Speech Recognition (ASR) in embedded appliances such as car-navigation systems, personal digital assistants and robots. In these appliances, processing power and available memory are generally restricted at a cost-conscious point of view, as not only ASR but also other applications are operating at the same platform. In such a case, the number of parameters contained in the acoustic model should be reduced. Then, the acoustic model cannot demonstrate enough performance even if it is trained with huge speech corpus covering variable conditions. Acoustic modeling specialized for the objective task is expected. To cope with differences in speaker characteristics and speaking style, speaker adaptation technique [1] [2] and speaker clustering technique [3] are proposed. For background noise, noise reduction technique [4], noise adaptation technique [5] and noise matched acoustic model are proposed and have been shown effective. In recent research [6], task dependency and reusability in four speech corpora were investigated by cross task recognition.

In this paper, the reusability of speech corpora is investigated. Therefore, six Japanese speech corpora is evaluated by two approaches. One is cross task recognition as a usual approach. The other one is a statistical multidimensional scaling (MDS) method called as COSMOS (COmprehensive Space Map of Objective Signal) method that visualizes the speech corpus within two or three dimensional space [7][8]. Visualization is known as an effective technique to grasp the multidimensional space which humans cannot understand easily. It is expected that to comprehend the relationship among multiple speech corpora by using visualization of their acoustic space is effective in order to analyze the reusability of a speech corpus.

In the next section, an overview of multiple speech corpora is described. In Section 3, a concept and a formulation of the COSMOS method are described. In Section 4, reusability of these corpora is investigated through cross task recognition and acoustic space visualization. Finally, a summary and an outlook on future work are given in Section 5.

## 2. Speech Corpus

In this section, an overview of the six Japanese speech corpora evaluated in this paper is described. Each speech corpus consists of six categories:

1) Car navigation command, called *COM*
2) City name, called *CITY*
3) Person name, called *NAME*
4) Word of foreign origin, called *FWORD*
5) Continuous 4 digits, called *4DIGIT*
6) Kana (defined as the basic unit of Japanese pronunciation; mostly coinciding with a syllable), called *KANA*

These speech corpora are task specific speech data used in a car navigation system and are composed of isolated words and short sentences. The size of each corpus is shown in Table 1.

Table 1 *Corpus Size*

| | Speakers | Utterances per speaker | Total size [h] |
|---|---|---|---|
| *COM* | 67 females | 472 | 9.8 |
| *CITY* | 76 females | 433 | 20.9 |
| *NAME* | 85 females | 256 | 8.9 |
| *FWORD* | 84 females | 351 | 11.2 |
| *4DIGIT* | 72 females | 140 | 4.9 |
| *KANA* | 75 females | 220 | 3.7 |

## 3. Statistical MDS

### 3.1. Concept of Visualization

In statistical pattern recognition such as speech recognition, training corpus has one of the biggest impacts on recognition performance. Therefore, building a training corpus requires

much attention. However,as it is difficult to grasp the training corpus consisting of multidimensional data such as speech, the common approach is to build it experimentally and evaluate it by only recognition performance. The multidimensional scaling (MDS) method [9] featuring a visual mapping of multidimensional information onto a visible space of low order (one to three) dimension is extremely effective in enhancing perceptibility of multidimensional data space such as acoustic space expressed with speech data.. The drawback of this method, however, is quickly growing computational complexity as the amount of data increases. On very large data sets, its applicability can become uncertain.

### 3.2. COSMOS method

The proposed method called COSMOS (COmprehensive Space Map of Objective Signal) [7][8] handles statistical model, such as GMM and HMM, as an approximated expression of the multidimensional data space representing the corpus. The method performs a nonlinear projection of acoustic models represented by a set of HMMs into visible dimensional space and was proposed as an extension of the Sammon method [10]. The Sammon method is carried out based on the mutual distances between vectors in multidimensional space, and does not depend on class labels or other discriminative information (unsupervised approach). As an improvement, by defining a distance between statistical models, it can be extended to map these statistic modelsminimizing projection-error from multiple to two dimensions. Although the PCA method may be applied to statistic model mapping by using super vectors concatenating only the mean values of Gaussian distributions, it was suggested that the information loss in mapping is big due to rather small cumulative proportion up to the second principle component [11].

#### *3.2.1. Formulation*

First, an overview of the Sammon method is described. In the Sammon method, the error function $E_m$ in formula (1) is minimized iteratively by the steepest descent method. $D(i,j)$ denotes mutual distance between the vector $i$ and $j$ existing in higher order dimensional space. $D_m(i,j)$ denotes mutual Euclidean distances of the mapped lower order coordinates of the vector $i$ and $j$ at $m$ th iteration. Generally, initial position mapped onto lower dimension is initialized by random value. And initial mutual distance $D_0(i,j)$ is computed from the position.

$$E_m \equiv \frac{1}{c}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left[\{D(i,j)-D_m(i,j)\}^2 / D(i,j)\right] \quad (1)$$

$$c \equiv \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} D(i,j) \quad (2)$$

Second, the extension of the Sammon method to a statistic model mapping is described. In this paper, a statistic model is an acoustic model based on HMM. In general, an acoustic model is a generic designation for an aggregation consisting of multiple models of acoustic units. Accordingly, the mutual distance $D(i,j)$ between acoustic model $i$ and $j$ is defined by the following:

$$D(i,j) \equiv \sum_{k=1}^{K} d(i,j,k) * w(k) \Big/ \sum_{k=1}^{K} w(k) \quad (3)$$

Here, $d(i,j,k)$ denotes the mutual distance between the acoustic unit $k$ within the acoustic model $i$ and the acoustic unit $k$ within the acoustic model $j$. $w(k)$ represents weight value such as an occurrence frequency for the acoustic unit $k$. $K$ is the total number of acoustic units. Assuming all acoustic models ($i = 1,...,N$) share a common topology with one-on-one state alignment between respective acoustic models, $d(i,j,k)$ can be defined using formula (4).

$$d(i,j,k) \equiv \frac{1}{S(k)}\sum_{s=0}^{S(k)-1}\frac{1}{L}\sum_{l=0}^{L-1} dd(i,j,k,s,l) \quad (4)$$

$S(k)$ represents the number of states in acoustic unit $k$. $L$ stands for the dimension of acoustic feature parameters. The Bhattacharyya distance [12] and Kullback Leibler distance [13] are commonly used as the distance measure $dd(i,j,k,s,l)$ between Gaussian distributions. In this paper, the Bhattacharyya distance is adopted.

The resulting visualized map itself is called COSMOS map, while the proposed mapping method is referred to as COSMOS method. Each single acoustic model projected onto the COSMOS map is called a STAR.

## 4. Analysis

### 4.1. Cross Task Recognition

Table 2 shows the conditions of this experiment. Evaluation speakers are randomly selected from each speech corpus shown in Table 1. Except evaluation data, all speech data is used for training data. All speech samples are overlaid with the background noise recorded at an exhibition hall and with a Signal-to-Noise ratio of 20 dB. Sampling frequency is 11.025kHz. A specific language model is constructed for each of the six tasks. During recognition, the task specific language models are used corresponding to that task'sevaluation data. Every task specific acoustic models ( named after its task) is trained with task specific training speech data. In addition, a task independent acoustic model (called as *TI*) is trained with all training data of the six tasks. Furthermore, for each task, an acoustic model (called as *woTI*) is trained with all training speech data, but without the task specific speech data. For instance, for evaluating the evaluation data in speech corpus *COM*, the *woTS* is trained with training speech data of *CITY*, *NAME*, *FWORD*, *4DIGIT* and *KANA*. In this experiment, an acoustic model structure is a mono-phone HMM expressed by single Gaussian distribution. The number of mono-phones is 29, and the number of states in each HMM is 3. The acoustic feature parameters consist of 10 MFCCs, 10 delta-MFCCs and 1 delta-log power.

Table 3 summarizes the cross task recognition performance. In each evaluation task, the difficulty of speech recognition depends on each language model size or complexity. For instance, the evaluation task *CITY* is an easy task because utterances are much longer than that of other categories.

Table 2 *Conditions of Experiment*

| *Task* | *Evaluation speakers* | *Language model size* |
|---|---|---|
| *COM* | 12 females | 472 words * 1 loop |
| *CITY* | 12 females | 3428 words * 1 loop |
| *NAME* | 12 females | 256 words * 1 loop |
| *FWORD* | 12 females | 501 words * 1 loop |
| *4DIGIT* | 12 females | 10 digits * 4 loops |
| *KANA* | 12 females | 110 words * 1 loop |

Table 3 *Cross Task Recognition Performance (Average of Word Accuracy [%])*

| Model | Evaluation data (female) | | | | | |
|---|---|---|---|---|---|---|
| | *COM* | *CITY* | *NAME* | *FWORD* | *4DIGIT* | *KANA* |
| *COM* | **82.6** | 96.9 | 92.4 | 89.1 | 97.0 | 37.1 |
| *CITY* | 78.1 | **98.6** | 93.9 | 85.3 | 95.9 | 37.0 |
| *NAME* | 73.7 | 97.9 | **95.2** | 84.7 | 96.4 | 38.1 |
| *FWORD* | 78.4 | 97.0 | 90.4 | **93.8** | 97.5 | 40.2 |
| *4DIGIT* | 73.0 | 91.8 | 83.9 | 82.9 | **99.0** | 35.0 |
| *KANA* | 66.2 | 73.7 | 74.8 | 74.1 | 84.6 | **53.8** |
| *TI* | 82.3 | 98.5 | 94.5 | 90.6 | 97.8 | 43.9 |
| *woTS* | 81.2 | 97.8 | 93.6 | 88.3 | 97.1 | 42.6 |

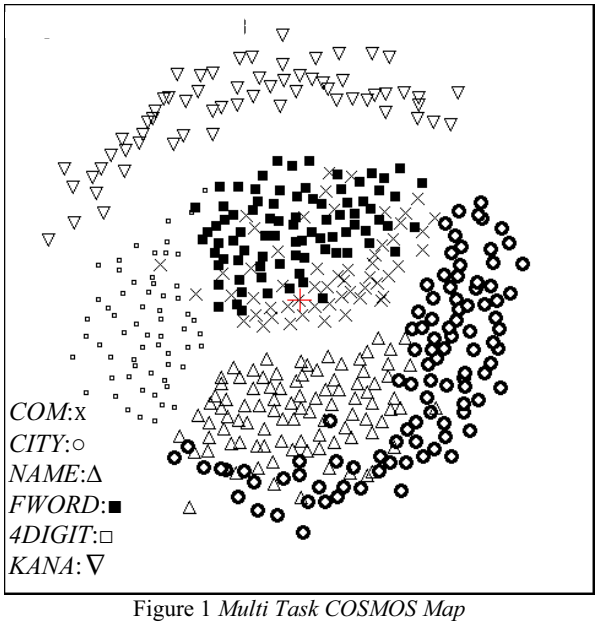


Figure 1 *Multi Task COSMOS Map*

### 4.2. Visualization

Visualization of multiple corpora is done using the COSMOS method. Initially, task dependent and speaker dependent acoustic models (called as *TSD*) are trained with each training speaker's speech data. The acoustic model structure is the same as of the models, which are defined in section 4.1. Then, each *TSD* is mapped by using the COSMOS method. Although a training speech data size for each *TSD* is different, an influence of a phoneme $k$ trained with few speech data is reduced by a weight $w(k)$ in formula (3). Figure 1 shows the COSMOS map constructed with 458 female *TSD*s. One *TSD* corresponds to one STAR (point) on the COSMOS map. In Figure 1, all *TSD*s sharing the same task are indicated by the same symbol. This map is considered to be an approximation of the whole acoustic space covering the multiple corpora in Table 1.

### 4.3. Discussion

#### *4.3.1. Analysis of Corpus Reusability*

First, the cross task recognition performance shown in Table 3 is discussed below. The highest performance is obtained when the task of acoustic model and evaluation data is the same, which is transcribed in a bold. The second highest performance (excluding performance of *TI* and *woTS*) is underlined. *TI* and *woTS* show robust performance for each evaluation task. It is expected that these acoustic models are effective for a system including several tasks. As to the performance of each task specific acoustic model, it can be said that it is degraded in different task situation. The performance gap depends on each cross task recognition pair and seems to represent the reusability of each speech corpus. Local relationship between two corpora, for instance, *CITY* and *NAME,* is shown clearly in this table. However, as the number of tasks evaluated in cross task recognition is increased, it is harder to judge the global relationships between multiple tasks by only taking into account numerical scores of cross task recognition performance.

On the other hand, an aggregate of each speech corpus is clearly separated in the COSMOS map shown in Figure 1. It suggests that the reusability of each speech corpus is not high enough. In particular, reusability of *KANA* seems much lower since it is located far from the other tasks with a big spatial gap on the COSMOS map. Exceptionally, the aggregate of *COM* and that of *FWORD* are overlapped closer. In the same way, aggregate of *CITY* and that of *NAME* are closer. It is expected that these two speech corpora would have high reusability between each other. In this way, a visualized map showing positions of speech corpora is effective to grasp their relationships.

Next, a correspondence between the cross task recognition performance shown in Table 3 and the position relationship shown in the COSMOS map of Figure 1 is described. In case of *COM* and *FWORD*, which are located closer on the COSMOS map, the performance gap between them is much smaller than the ones between *COM* and other tasks, and ones between *FWORD* and other tasks. *COM* and *FWORD* are complementary and reusable for each other, so are *CITY* and *NAME*. In addition, the performance of *woTS* is discussed as follows. An acoustic model, which is closer within the COSMOS map or gives second highest performance, shows higher performance than *woTS* except *COM* and *KANA* cases. In *COM* case, an acoustic space of *COM* located at the center on the COSMOS map is interpolated with other tasks surrounding *COM*. It is suggested that the robust performance of *woTS* to *COM* accounts for this interpolation. On the other hand, in case of *KANA* located far from other tasks, the performance gap between *KANA* and other tasks is much bigger. This difference in performance corresponds to a big spatial gap on the COSMOS map. Reusability cannot be expected with five other tasks for *KANA*. In order to improve the recognition performance of *KANA*, it seems necessary to collect more speech data of *KANA* itself.

These facts indicate that the local relationship of position on the COSMOS map seems almost corresponding to the cross task recognition performance.

#### 4.3.2. *Key Issue of Task Dependency*

Figure 2 shows the occurrence frequency rate of twelve highly frequent mono-phones in each speech corpus in Table 1. Figure 3 shows the map, which is visualized by the Sammon method, of occurrence frequency rate vector. The position of each speech corpus in Figure 1 and the one in Figure 3 are correlated. It is suggested that the occurrence frequency rate of a mono-phone in each speech corpus is one of the key issues of task dependency.

Thus, the COSMOS map constructed from multiple speech corpora represents the global relationship of multiple tasks and suggests reusability of each speech corpus to other tasks. We can expect that any speech corpora, which are located closer to the corpora of the target task on the COSMOS map, are reusable to enhance acoustic model for ASR in the target application, even though they are collected for completely different applications. This is a new important point of view.

## 5. Summary and Future Work

In this paper, the new technique of analyzing reusability of speech corpus by using a statistical multidimensional scaling method is proposed. It was demonstrated that the visualized map of multiple speech corpora showed quite high correspondence to cross task recognition performance score. In addition, the global relationship between multiple speech corpora was clearly represented on the visualized map.

In our future work, a new technique of how to reuse large volume of available multiple speech corpora, which have been already collected by many organizations in last decade, will be investigated. In addition, how to cover a spatial gap and sparseness in the COSMOS map by utilizing other task speech corpora will be investigated.

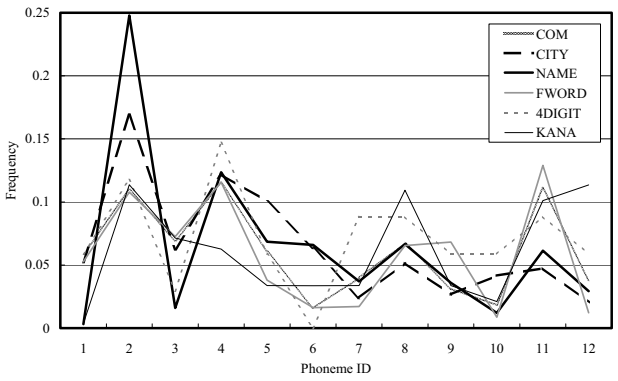


Figure 2 *Occurrence Frequency Rate of Mono-phone in Each Speech Corpus*

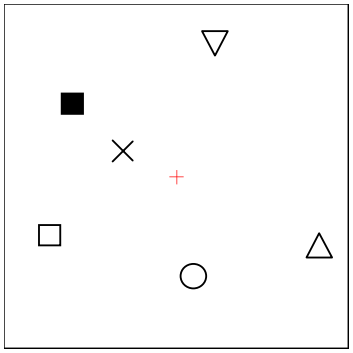


Figure 3 *Occurrence Frequency Rate Vector Map*

## 6. References

[1] C. J. Leggretter et al., "Maximum likelihood linear regression for speaker adaptation of continuous-density hidden Markov models," Computer Speech and Language, vol.9, pp.171-185, 1995.

[2] J. L. Gauvain et al., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. Speech Audio Process., vol.2, no.2, pp.291-298, 1994.

[3] T.Kosaka et al., "Tree-structured speaker clustering for speaker-independent continuous speech recognition," ICSLP-94, pp.1375-1378, 1994.

[4] M. Shozakai, S. Nakamura and K. Shikano, "A speech enhancement approach E-CMN/CSS for speech recognition in car environments," Proc. IEEE ASRU-97 Workshop, Santa Barbara, USA, pp.450-457, 1997.

[5] M. J. F. Gales, et al. "An improved approach to the hidden Markov model decomposition of speech and noise," Proc. ICASSP92, pp.233-236, 1992.

[6] F. Lefeve et al., "Genericity and portability for tash-independent speech recognition," Computer speech and language 19, pp.345-363, 2005.

[7] M. Shozakai et al., "Acoustic space analysis method utilizing statistical multidimensional scaling technique," Proc. NSIP-05, Sapporo, Japan, May 2005.

[8] G. Nagino et al., "Building an effective corpus by using acoustic space visualization (COSMOS) method," IEEE ICASSP, vol. I, pp.449-452, 2005.

[9] A. K. Jain et al., "Statistical pattern recognition: a review," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp.4-37, 2000.

[10] J. W. Sammon, "A nonlinear mapping for data structure analysis," IEEE Trans. Computers, vol.C-18, no.5, pp.401-409, May 1969.

[11] A. Nagorski et al., "Optimal selection of speech data for automatic speech recognition system," Proc. ICSLP, vol.4, pp.2473-2476, 2002.

[12] K. Fukunaga, "Introduction to statistical pattern recognition (Second edition)," Academic Press, Inc., San Diego, 1990.

[13] H. Akaike, "A new look at the statistical model identification," IEEE Trans. Autom. Control, vol.16, no.6, pp.716–723, Dec. 1974.